脳を作るプロジェクト
Make Brain Project

太田怜志 †　Reiji Ota

†School of Systems Information Science, Future University Hakodate
116-2, Kamedanakano-cho, Hakodate-shi, Hokkaido 041-0803, Japan
Email: b1021121@fun.ac.jp

1. 背景

日々進歩を遂げる人工知能の研究は、我々の生活のあらゆる側面に影響を与えている。その成果は、我々が日常的に利用する多くの便利なツールとなり、生活の質を向上させている。しかし、人工知能は多くの分野で活用されているが、計算コストの膨大さなどの理由から、まだ全ての分野で十分に活用されているわけではない。そこで本プロジェクトは、脳の仕組みを取り入れた低コストな人工知能の開発や、現実問題に人工知能を応用することを目的として活動した。

本プロジェクトは、音楽生成班、自動運転班、動画要約班の 3 つの班に分かれて活動を行った。音楽生成班は、脳の仕組みを取り入れた、音楽生成 AI の開発を試みた。自動運転班は、カメラ入力のみで自動運転を実現させることを試みた。動画要約班は、動画の重要な箇所だけを切り抜いた動画を作成するシステムの開発を試みた。

2. 課題の設定と到達目標

2.1. 音楽生成班

近年、生成 AI が多くの分野で注目を集めている。しかし、生成 AI の課題の 1 つとして、学習コストが膨大であることが挙げられる。生成 AI は大量のデータを必要とし、そのデータを学習するために高性能なコンピュータを使い、学習の過程で大量の電力を消費する。音楽生成班では、脳の仕組みを取り入れた音楽生成 AI を開発することで、学習コストの削減を試みた。

2.2. 自動運転班

近年、世界的に自動運転技術の研究が進展している。従来の自動運転車 [1] は高精度なセンサーやレーダーを搭載し、自動運転を実現している。しかし、これらのセンサーやレーダーは高価格であることが課題として挙げられる。そこで、自動運転班では、安価なカメラを用いて、カメラ入力のみで、自動運転を達成する方法の開発を試みた。

2.3. 動画要約班

昨今、動画の倍速視聴の利用が増えている。しかし、倍速視聴では内容が理解できないなどの懸念点が挙げられている [2]。そこで、動画要約班では、動画を短く編集し、全体像を短時間で把握出来るように要約するシステムの開発にとり組んだ。

3. 課題解決のプロセスとその結果

3.1. 音楽生成班

音楽生成班は音楽生成 AI の開発を 2 つの異なるアプローチで試みた。1 つ目は、テキストから音楽を生成するアプローチ（以下「T2M アプローチ」とよぶ）である。2 つ目は、ある音楽を別のジャンルの音楽に変更するアプローチ（以下「M2M アプローチ」とよぶ）である。2 つのアプローチで開発を試みた理由は、リスクの分散と開発の効率化である。2 つのアプローチで並行して開発を進めることで、それぞれの収集したデータや得られた知見を共有することが可能となる。これにより、同じ失敗を繰り返すことを防ぎ、開発の効率化ができると考えた。

3.1.1. T2M アプローチ

T2M アプローチでは、Stable Diffusion[3] を用いて、テキストから音楽のスペクトログラムを生成することを試みた。スペクトログラムとは、音の周波数成分と時間変化を色で表示するグラフである。このアプローチでは Stable Diffusion に対して、Fully Spiking Variational Autoencoder[4]（以下「FSVAE」とよぶ）を導入し、学習コストの削減を試みた。

T2M アプローチでは、既存の Stable Diffusion を改善することでテキストから音楽を生成することを試みた。Stable Diffusion とは、テキストから画像を生成するモデルである。Stable Diffusion は、Autoencoder、U-NET と Text Encoder によって構成されている。Autoencoder は、画像と特徴量を相互に変換する役割を持つ。Autoencoder により画像の基本的な意味を捉えることが期待される。U-NET とは、畳み込みニューラルネットワークの一種であり、画像に付与されたノイズを学習するものである。Text Encoder は、テキストをベクトルに変換する。Stable Diffusion では次の手順でテキストから画像を生成する。まず、画像を Autoencoder に入力し、拡散プロセスでノイズを連続的に付与する。次に、テキストを Text Encoder に入力し、出力を得る。得た出力と、ノイズが付与された画像を一度に U-NET に入力し、逆拡散プロセスでノイズを除去する。最後にノイズが除去された画像を Autoencoder に入力し、画像を生成する。

T2M アプローチでは、既存の Stable Diffusion に対して、以下の 2 つの改善を行い、学習コストの削減を試みた。1 つ目は、Autoencoder を FSVAE に置き換えることである。FSVAE とは、Autoencoder にスパイキングニューラルネットワークを用いたものである。スパイキングニューラルネットワークとは、ニューロンの膜電位をモデル化したニューラルネットワークである。スパイキングニューラルネットワークで扱う数値は、実数値ではなく、0 または 1 の 2 値となる。2 つ目は、U-NET 内に Token-shift[5] を組み込むことである。Token-shift とは、Token と呼ばれる、テキストを意味単位で分割したものを、わずかに移動させる機構である。Token-shift を利用することで、言語処理においては、計算量の削減に成功している。

T2M アプローチの提案モデルは、次の手順でテキストから画像を生成する。まず、音楽をスペクトログラムの画像に変換する。スペクトログラムをコピーして重ね、FSVAE に入力し、特徴量を抽出する。抽出した特徴量について、拡散プロセスにて、ガウス分布に従うノイズを連続的に付与する。続いて、テキストを Text Encoder に入力し、出力を得る。得られた出力と、ノイズが付与されたスペクトログラムの特徴量を一度に、Token-shift を組み込んだ U-NET に入力し、逆拡散プロセスでスペクトログラムの特徴量におけるノイズを連続的に除去する。続いてノイズが除去された特徴量を、FSVAE に入力し、連なったスペクトログラムを生成する。最後に、連なったスペクトログラムを足し合わせ、逆フーリエ変換を行い、音楽に変換する。

結果として、生成した音楽はノイズを含んでいた。また、FSVAE を用いたことで、既存のモデルと比べて、大量のメモリを消費していた。これは学習コストを削減するという目的に反している。しかし、既存のモデルと比べて提案モデルは、学習の収束が早いことが分かった。

3.1.2. M2M アプローチ

M2M アプローチでは、CycleGAN を用いて、ある音楽を別のジャンルの音楽に変更することを試みた。CycleGAN は敵対的生成ネットワーク（以下「GAN」とよぶ）の一種で、ある音楽のジャンルから別の音楽のジャンルへの変換を学習することができる。このアプローチでは、CycleGAN に対して、レザバー計算を導入し、学習コストの削減を試みた。

はじめに、GAN を用いた音楽生成に取り組んだ。GAN は Generator と Discriminator の二つのネットワークから構成されるディープニューラルネットワークの一種である。Generator はデータを生成し、Discriminator はそのデータが本物か偽物かを判別する。Generator は Discriminator を欺こうと学習し、Discriminator はより正確に識別しようと学習する。これにより、GAN は本物に近いデータを生成する能力を向上させることが期待される。しかし生成した音楽は高音が鳴るだけで、リズムやメロディーが生成できていなかった。生成した音楽のスペクトログラムを見てみると図 1 に示すような周期的なパターンが出現していることがわかった。調査を進めると Discriminator が意図しない学習を行っていることがわかった。

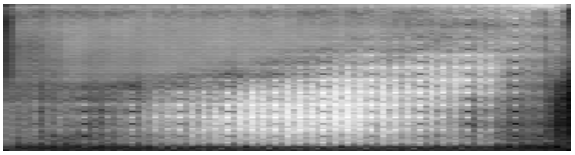

図 1: GAN を用いて生成した音楽のスペクトログラム

次にこの問題を解決するために、WaveGAN を導入した。WaveGAN は GAN の Discriminator に Phase Shuffle という操作を行う層を導入したものである。Phase Shuffle を行うことで、Discriminator の意図しない学習を防ぐことができる。生成した音楽に音の乱れはあったが、リズムやメロディーを生成することができた。

さらに、WaveGAN の学習コスト削減を実現するため、Generator にレザバー計算を導入する。レザバー計算は時系列の学習が可能なリカレントニューラルネットワークの枠組みである。レザバー計算は、入力層、レザバー層、出力層から構成される。レザバー計算におけるネットワークの学習は、レザバー層から出力層への結合に限定しており、学習コストが少ない。今回は、レザバー計

算の代表的なモデルであるエコーステートネットワーク [6]（以下「ESN」とよぶ）を用いた。生成した音楽はメロディーは不十分だったが、リズムは生成できていた。また 1 エポックあたりの学習時間を約 20 秒短くすることができ、学習コストの削減に成功した。メロディーも生成できるよう調査したが、望ましい結果が得られず、CycleGAN を用いた音楽のジャンル変換まで実現させることができなかった。

3.2. 自動運転班

自動運転班では、カメラ入力のみで信号機を含む交差点で右左折ができれば、自動運転が実現できるのではないかと考えた。そこで自動運転班では、自動運転の実現を 3 段階に分けて実現することを決定した。1 段階目は直進・カーブの実現、2 段階目は道路標識の認識の実現、3 段階目は信号の認識の実現である。以上の段階を実現するために、自動運転班ではシミュレーション環境を活用した。シミュレーション環境を Unity を用いて作成し、各段階での自動運転の性能をテストした。

自動運転班では、強化学習を用いて、自動運転車の学習を試みた。強化学習とは、ある環境内におけるエージェントが、試行錯誤しながら最適な行動を学習する、機械学習の枠組みの一つである。自動運転車の強化学習におけるエージェントとは自動運転車の車両であり、環境とは道路標識、信号などを含む、道路状況である。今回は強化学習の中でも Dueling Network を組み込んだ Double Deep Q-Networks[7]（以下「Dueling DDQN」とよぶ）を採用した。Dueling DDQN を採用した理由は、強化学習アルゴリズムの中でも、高速かつ安定した学習の収束が可能だからである。

3.2.1. 1 段階目：直進・カーブの実現

1 段階目では、次の手順で学習を行った。まず、単眼カメラを用いて、自動運転者の進行方向の画像を取得する。次に、その取得した画像から、車線を検出する。この検出作業により、左右の車線の交点が明らかになり、これを消失点と定義する。その後、消失点が中心から左右どちらに傾いているかを数値化する。この数値を強化学習の環境とする。自動運転車はこの環境に基づいて、どのように動くべきかを判断する。最後に、Dueling DDQN を用いて、この環境内の自動運転車の車両が車線を逸脱しないように学習を行う。

結果として、直進・カーブのシミュレーションは、学習の途中段階である。このシミュレーションにおいて、期待した結果を得ることができなかった。

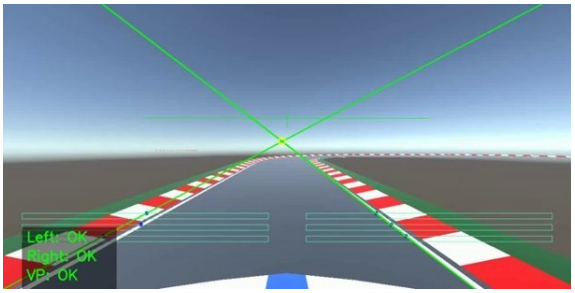

図 2: シミュレーション環境での学習の様子

3.2.2. 2 段階目：道路標識の認識

2 段階目では、画像処理のアルゴリズムが実装できず、このシミュレーションは実装には至らなかった。

3.2.3. 3 段階目：信号の認識

3 段階目も 2 段階目と同様に、このシミュレーションは、時間が足りず、実装には至らなかった。

3.3. 動画要約班

動画要約班では、動画の全体像を短時間で把握できるように要約するシステムの開発を行った。はじめに、動画の要約に必要な手法の調査を行った。ここで、画像を生成する技術、音声をテキストに変換する技術、テキストから重要なキーワードを抽出する技術などを検討した。次にデータセットの作成を行った。函館の観光地を撮影環境が異なるように撮影した。撮影したデータにラベルやメタデータなどの情報を付与し、学習データとした。作成したデータセットを用いて物体検出モデルの学習を行った。YOLO[8] などのアルゴリズムを適用し、データ内の特徴的な物体やテキストを検出できるように学習を行った。最後に、次の手順で動画を短くする。画像認識 [8] を用いて動画内にある特徴物を検出・トリミングを行い、指定された大きさに加工する。その画像を４枚組み合わせて一枚にし、４枚が揃うまでの時間に発せられた言葉を要約して中央に表示する。その１枚の画像を５秒間表示する。以上の手順を繰り返し、動画を作成する。

結果として、当初の目的であった、動画の作成に成功した。しかし、画像認識のため学習させた物体以外をトリミングすることができなかった。また、要約した動画について以下の手順でアンケートを行った。はじめに、被験者に、要約前の動画を見てもらう。そして、要約前の動画についての感想と、写されたオブジェクトを可能な限り多く書き出してもらう。次に、被験者に、要約した動画を見てもらう。そして、要約した動画に対して、

感想を書いてもらう。以上のアンケートを行ったところ、一度に情報をすべて把握することが難しいという感想が見られた。

4. 今後の課題

4.1. 音楽生成班

4.1.1. T2M アプローチ

T2M アプローチはテキストから音楽の生成を実現させることができなかった。課題として、大量のメモリを消費していたことが挙げられる。今後は、学習データとしてスペクトログラムを扱う手法ではなく、波形データを用いる手法や MIDI データを用いるなどの改善が必要である。

4.1.2. M2M アプローチ

M2M アプローチは音楽のジャンル変換を実現させることができなかった。課題として、生成した音楽にノイズが含まれていたことが挙げられる。原因としては、学習データが不足であった点とモデルのパラメータが適切ではなかった点の 2 点が考えられる。今後は、学習データをより充実させ、パラメータ探索を行う必要がある。

4.2. 自動運転班

自動運転班で行ったシミュレーションでは、カメラ入力のみで信号機を含む交差点で右左折が可能であることを示すことはできなかった。期待した結果を得られなかった原因として、シミュレーションに時間がかかってしまったことが挙げられる。シミュレーションに時間がかかることで、適切なパラメータを用いることができなかった。今後の課題は、シミュレーションを高速化し、直進・カーブの精度を向上させることである。

4.3. 動画要約班

今後の課題は、編集した画像の出力方法を変更することである。評価アンケートにおいて、視聴するのが大変という意見が見られた。現在はトリミングをした画像を 4 枚まとめて出力し、音声を要約した文章を中央に配置している。そのため、出力の形式を見やすいものに変更する必要がある。

参考文献

[1] 日本経済新聞, “日立系、単眼カメラだけの低コスト自動運転センサー”, https://www.nikkei.com/article/DGXZQOUC124340S3A110C2000000/ (2024/01/09 最終アクセス)

[2] Cross Marketing, “動画の倍速視聴に関する調査（2021 年）”, https://www.cross-m.co.jp/report/life/20210310baisoku/ (2024/12/20 最終アクセス)

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models” *CVPR*, 2022.

[4] Hiromichi Kamata, Yusuke Mukuta, and Tatsuya Harada, “Fully Spiking Variational Autoencoder.” *AAAI*, 2022.

[5] Hao Zhang, Yanbin Hao, Chong-Wah Ngo, “Token Shift Transformer for Video Classification.” *MM’21*, pp.917–925, 2021.

[6] H. Jaeger, “A tutorial on training reccurent neural networks, covering BPPT, RTRL, EKF and ”echo state network” approach,” *GMD Report*, vol.159, pp.1–46, 2002.

[7] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas, “Dueling Network Architectures for Deep Reinforcement Learning,” *PMLR*, vol.48, pp.1995–2003, 2016.

[8] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, Bo Ma, “A Review of Yolo Algorithm Developments,” *Procedia Computer Science*, vol.199, pp.1066–1073, 2022.